Die Behandlung fehlender Werte in logischen Ausdrücken bei SAS und SPSS: eine Warnung vor unerwarteten Ergebnissen

Ritter, Heiner; Züll, Cornelia; Grüner, Hans

Die Behandlung fehlender Werte in logischen Ausdrücken bei SAS und SPSS: Eine Warnung vor unerwarteten Ergebnissen

Bei der Fallauswahl (Kommandos "IF" bei SAS bzw. "IF" und "SELECT IF" bei SPSS) kommen SAS, SAS/PC und SPSS 9 im Gegensatz zu SPSS-X und schließlich SPSS/PC+ zu unterschiedlichen Ergebnissen, also zu unterschiedlich ausgewählten Fallzahlen. Dies hat seine Ursache in der unterschiedlichen Behandlung von fehlenden Werten (Missing Data) in logischen Ausdrücken. Im Beitrag werden anhand der systematisch vergleichenden Darstellung sowie an einem praktisch durchgerechneten Beispiel die Auswirkungen dargestellt.

1. Vorbemerkung

In den ZUMA-Nachrichten 19 haben wir über einige Ergebnisse von Vergleichstests zwischen den PC- und Mainframe-Versionen der Statistik-Programmsysteme SAS und SPSS berichtet. Bei diesem Vergleich wurde vor allem Wert gelegt auf den Leistungsumfang der PC-Version gegenüber der jeweiligen Mainframe-Version und auf die Handhabbarkeit der PC-Versionen für den Benutzer. Nicht näher berücksichtigt wurde jedoch die Frage nach der Funktionsweise gleicher oder ähnlicher Kommandos für das Datenmanagement in den Programmsystemen. Wichtig ist dabei in erster Linie die gleiche Funktionsweise zwischen einer alten und neuen Mainframe-Version des Programmpakets bzw. der entsprechenden PC-Version. Die meisten Vergleiche von Programmsystemen haben hauptsächlich die Güte und Vergleichbarkeit von Statistikprozeduren zum Thema. Der Bereich der Datenmodifikation und -selektion und dabei auch die Behandlung von fehlenden Werten wird bei diesen Vergleichen und Überprüfungen zumeist ausgeklammert. Er findet oft nur insoweit Beachtung, wie es um erweiterte Möglichkeiten bei der Schleifenbildung, bei Verkürzungen, bei Sprüngen, etc. geht. Was die implementierte Logik angeht, so wird höchstens darauf geachtet, daß logische Operatoren vorhanden und komplexe logische Ausdrücke (Verknüpfungen) möglich sind. Die Frage aber, ob z.B. gleiche Datenselektionskommandos gleiche Ergebnisse liefern oder nicht, und wie dabei fehlende Werte behandelt werden, wird vernachlässigt.

Auf ein mögliches Problem beim Arbeiten mit logischen Abfragen muß der Anwender beim Umsteigen von SPSS 9 auf SPSS-X oder SPSS/PC+ aufmerksam gemacht werden. Er stellt zu seiner Überraschung - meist mehr oder weniger zufällig - fest, daß seine "IF" oder "SELECT IF" Abfragen andere Ergebnisse liefern als in SPSS 9. Dasselbe gilt auch für Anwender, die von SPSS-X auf SPSS/PC+ (oder umgekehrt) umwechseln. Auf Unterschiede bei der Behandlung von fehlenden Werten in logischen Ausdrücken wird in den SPSS-Handbüchern unter "Help for Old Friends" (Mainframe) und "Help for SPSS-X Users" (PC) zwar formal korrekt hingewiesen, aber aus unserer praktischen Erfahrung mit Benutzerberatungen wissen wir, daß die Anwender in aller Regel solche Hinweise zunächst einmal ignorieren, da sie sich der Konsequenzen nicht bewußt sind.

Im folgenden versuchen wir darzustellen, welche Unterschiede für den Benutzer beim Arbeiten mit den Datenselektionskommandos auftreten, wenn seine Daten fehlende Werte (Missing Data) enthalten und welche Konsequenzen sie für seine Analysen haben können. Wir haben bei unserer Gegenüberstellung SAS (Release 5.16 (Mainframe), Release 6.02 (PC)), SPSS-X (Release 2.2) und SPSS/PC+ (Update Version vom März 1987) berücksichtigt. Zusätzlich haben wir SPSS 9 miteinbezogen, da viele SPSS-X- und SPSS/PC+-Anwender zunächst mit SPSS 9 gearbeitet haben, bevor sie auf die neuen Systeme umgestiegen sind.

2. Systematische Darstellung der Missing Data Behandlung in logischen Abfragen

Wie schon oben gesagt, behandeln die von uns ausgewählten Programmsysteme logische Ausdrücke, die fehlende Werte enthalten, auf ganz unterschiedliche Art und Weise. SAS und SAS/PC behandeln diese völlig identisch; wie ältere Versionen von SAS in diesem Fall gearbeitet haben, ist für uns nicht mehr nachvollziehbar. SPSS 9, SPSS-X und SPSS/PC+ dagegen handhaben die fehlenden Werte auf jeweils verschiedene Arten. SPSS 9 kennt nur Benutzer-definierte Missing Data, die anderen Systeme dagegen unterscheiden zwischen Benutzer-definierten und System-Missing-Data-Werten. Diese Systeme machen jedoch keinen Unterschied, ob es sich um System- oder um Benutzer-definierte Missing-Data-Werte handelt. Vereinfacht ausgedrückt bedeutet das:

- SAS, SAS/PC und SPSS 9 handhaben die Werte, als ob es sich um gültige Werte handeln würde.
- SPSS-X versucht, trotz fehlender Werte zu einem gültigen Ergebnis zu kommen. Nach der dort implementierten Logik kann das Ergebnis einer Verknüpfung zweier Bedingungen durch die logische Operation "AND" nie wahr "TRUE" sein, wenn eine der beiden Bedingungen unbestimmt ist (d.h. die abgefragte Variable enthält einen fehlenden Wert). Dagegen kann das Ergebnis, abhängig vom Ergebnis der anderen Bedingung, falsch "FALSE" werden. Sind zwei Ausdrücke mit der Funktion "OR" verbunden, kann das Ergebnis nie "FALSE" sein, wenn das Ergebnis eines Ausdrucks unbestimmt ist, d.h. wenn der abgefragte Wert als fehlend deklariert ist. Das Ergebnis ist jedoch dann wahr, wenn das Ergebnis eines Ausdrucks wahr ist, auch wenn der andere unbestimmt ist.
- SPSS/PC+ liefert für logische Abfragen, in denen eine der Bedingungen fehlende Werte enthält, immer das Ergebnis "fehlend".

In SPSS-X und SPSS/PC+ steht zusätzlich eine Option "VALUE" zur Verfügung, mit der die gleiche Behandlung wie in SPSS 9 bzw. SAS erreicht werden kann.

In den folgenden Tabellen ist die Missing-Data-Behandlung bei logischen Ausdrücken in den Programmsystemen SAS, SAS/PC, SPSS 9, SPSS-X und SPSS/PC+ dargestellt. Zwei Bedingungen werden durch die Operatoren "AND" (Tabelle 1) bzw. durch "OR" (Tabelle 2) verbunden. Jeder Ausdruck kann das Ergebnis wahr "TRUE", falsch "FALSE" oder unbestimmt "MISSING" haben. "MISSING" bedeutet dabei, daß der Wert, der abgeprüft wurde, ein als fehlend deklarierter Wert war.

Tabelle 1: Fehlende Werte in logischen Ausdrücken in Verbindung mit dem Operator "AND"

Ergebnis Bedingung 1	Ergebnis Bedingung 2	Ergebnis "Bedingung 1 AND Bedingung 2" bei ... SAS,SAS/PC, SPSS 9	SPSS-X	SPSS/PC+
TRUE	TRUE	TRUE	TRUE	TRUE
FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE
TRUE	MISSING	*	MISSING	MISSING
MISSING	MISSING	*	MISSING	MISSING
FALSE	MISSING	FALSE	FALSE	MISSING

* in der obigen Tabelle bedeutet, daß alle Missing-Data-Werte als gültige Werte behandelt werden. Das Ergebnis ist in den Fällen "TRUE", in denen die Bedingung trotz fehlenden Werts zutrifft. Ist zum Beispiel in SAS die folgende Abfrage vorgegeben:
IF (KIRCHG EQ 2 AND BERUFST NE 4)
und KIRCHG hat den Wert 2, BERUFST enthält den System-Missing-Data-Wert, dann ist das Ergebnis wahr "TRUE".

Tabelle 2: Fehlende Werte in logischen Ausdrücken in Verbindung mit dem Operator "OR"

Ergebnis Bedingung 1	Ergebnis Bedingung 2	Ergebnis "Bedingung 1 OR Bedingung 2" bei ... SAS,SAS/PC, SPSS 9	SPSS-X	SPSS/PC+
TRUE	TRUE	TRUE	TRUE	TRUE
FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	TRUE	TRUE	TRUE
TRUE	MISSING	TRUE	TRUE	MISSING
MISSING	MISSING	*	MISSING	MISSING
FALSE	MISSING	*	MISSING	MISSING

* (siehe oben, Tabelle 1)

Wie fehlende Werte in logischen Abfragen behandelt werden, ist in den folgenden Handbüchern beschrieben:

SAS Institute Inc., 1985: SAS User's Guide: Basics, Version 5 Edition. Cary, NC: SAS Institute Inc. (Kapitel 16: Missing Values in Logical Operations);
SAS Institute Inc., 1985: SAS Language Guide for Personal Computers, Version 6 Edition. Cary, NC: SAS Institute Inc. (Kapitel 12: Missing Values in Logical Operations);
SPSS Inc., 1986: SPSS-X-User's Guide, 2nd Edition. Chicago: McGraw Hill (8.19 Missing Values and Logical Operators);
W. Schubö, H.-M. Uehlinger, 1986: SPSS-X-Handbuch der Programmversion 2.2. Stuttgart: Gustav Fischer (5.4.2.3. Logische Operatoren);
SPSS Inc., 1986: SPSS/PC+ for IBM PC/XT/AT. Chicago: SPSS Inc. (B36, 4.22 Missing Values).

3. Praktische Auswirkungen der unterschiedlichen Behandlung der fehlenden Werte

Welche Konsequenzen ergeben sich aus den oben beschriebenen Unterschieden für den Anwender? Am Beispiel des ALLBUS 86 Datensatzes haben wir versucht, die Auswirkungen zu verdeutlichen. Wir haben dazu Variablen mit relativ hohen Anteilen an fehlenden Werten ausgewählt:

V317 (KIRCHG) - Wie oft gehen Sie im allgemeinen in die Kirche? (Kategorien 1-4 zusammengefaßt zu 1 (regelmäßig), 5-6=2 (so gut wie nie), 0,9=9 (fehlend))
V198 (BERUFST) - Berufliche Stellung (zusammengefaßt zu Hauptgruppen, 4 sind dabei die Arbeiter)
V93 (STRICKEN) - Können Sie stricken? (zusammengefaßt in 1,2=1 (ja), 3=2 (nein), 0,9=9 (fehlend))
V196 (GESCHL) - Geschlecht (1 (männlich), 2 (weiblich))

Gibt es einen Zusammenhang zwischen dem Geschlecht und der Fähigkeit zu stricken bei Leuten, die nicht oder selten in die Kirche gehen oder keine Arbeiter sind? Erreicht wird die Fallselektion mit dem Kommando

```
SELECT IF (KIRCHG EQ 2 OR BERUFST NE 4)  (SPSS)  oder
IF (KIRCHG EQ 2 OR BERUFST NE 4);        (SAS).
```

Im folgenden haben wir uns nun darauf beschränkt, die Zellenbesetzung der Kreuztabelle der Variablen GESCHL und STRICKEN zu vergleichen. Diese Kreuztabelle unterscheidet sich je nach Programmsystem deutlich. Tabelle 3 ist mit SPSS 9 bzw. SAS generiert, Tabelle 4 mit SPSS-X und Tabelle 5 mit SPSS/PC+.

Tabelle 3: SPSS 9 und SAS: Kreuztabelle GESCHL/STRICKEN nach vorheriger Selektion

		STRICKEN ja	nein	Total
GESCHL	männlich	117	1148	1265
	weiblich	1508	96	1604
	Total	1625	1244	2869

Tabelle 4: SPSS-X: Kreuztabelle GESCHL/STRICKEN nach vorheriger Selektion

		STRICKEN ja	nein	Total
GESCHL	männlich	99	895	994
	weiblich	828	55	883
	Total	927	950	1877

Tabelle 5: SPSS/PC+: Kreuztabelle GESCHL/STRICKEN nach vorheriger Selektion

		STRICKEN ja	nein	Total
GESCHL	männlich	63	606	669
	weiblich	430	28	458
	Total	493	634	1127

Aus diesen Tabellen sieht man, welche Folgen die unterschiedliche Behandlung der Datenselektionen haben kann. Obwohl die Daten und die Analysestrategie jeweils völlig gleich sind, erhält man abhängig von der Programmversion

deutlich verschiedene Ergebnisse und kommt damit zu verschiedenen bzw. falschen Interpretationen.

Für den Anwender, der während der Analyse seiner Daten aus welchen Gründen auch immer das System wechselt, stellt sich damit die Frage, wie er Vergleichbarkeit erhalten kann. Für alle, die von SPSS 9 auf SPSS-X bzw. auf SPSS/PC+ wechseln, bleibt die Möglichkeit, Benutzer-definierte Missing-Data-Werte mit der "VALUE"-Funktion zu bearbeiten, um so dasselbe Ergebnis wie vorher in SPSS 9 zu erhalten. Die Abfrage müßte dann

```
SELECT IF (VALUE(KIRCHG) EQ 2 OR VALUE(BERUFST) NE 4)
```

lauten. Die Kreuztabelle wäre dann auch in SPSS-X und SPSS/PC+ dieselbe wie Tabelle 3. Will man in SPSS-X dieselben Ergebnisse wie in SPSS/PC+ erreichen, so müßte das obige Beispiel wie folgt formuliert werden:

```
SELECT IF ((NOT MISSING(KIRCHG) AND NOT MISSING(BERUFST)) AND (KIRCHG EQ 2 OR BERUFST NE 4))
```

Hier zeigt sich schon, daß es sehr komplex wird, in SPSS-X und SPSS/PC+ gleiche Ergebnisse erreichen zu wollen. Soll in SPSS/PC+ das gleiche Ergebnis wie in SPSS-X erreicht werden, so wird die Formulierung noch komplizierter und kann nur noch mit sehr komplexen "IF"-Abfrage, mit Hilfsvariablen und vorherigem Abprüfen auf fehlende Werte erreicht werden. Die Abfrage ist dann aber so undurchsichtig, daß die eigentliche Fragestellung nicht mehr erkennbar ist.

4. Schlußbemerkung

Daß bei gleichem Aufbau des Datenselektionsteils, abhängig vom ausgewählten System, unterschiedliche Ergebnisse berechnet werden, ist beunruhigend. Der Benutzer kann nur eindringlich darauf aufmerksam gemacht werden, sich genau zu vergewissern, ob er auch das Ergebnis, die Fallauswahl, so erhalten hat, wie er sie wirklich will. Beim Wechsel auf ein anderes System (z.B. von SAS auf SPSS-X oder umgekehrt) ist man in der Regel vorsichtig und achtet genauer auf Hinweise im Manual bzw. überprüft die Ergebnisse. Beim Wechsel von einer älteren Version auf eine neue Programmversion oder von der Mainframe-Version auf die PC-Version (oder auch umgekehrt) desselben Herstellers dagegen wird oft nicht mit der gleichen Vorsicht vorgegangen. Aber wie das obige Beispiel zeigt, ist auch hier besondere Vorsicht geboten.

Der Beitrag wurde von Heiner Ritter und Cornelia Züll in Zusammenarbeit mit Hans Grüner, Freie Universität Berlin, verfaßt.